1

Základní principy statistického popisu vícerozměrných dat, bodové a intervalové odhady základních charakteristik náhodných veličin.

KMA/MSM

- **vícerozměrná data** – obsahují několik atributů (dimenzí), tvoří body v nD prostoru
  - **vícerozměrné statistické metody**:
    1) **popis a analýza struktury vícerozměrných dat** – základní statistické zpracování (charakteristiky polohy a variability, outliers, transformace dat), grafická interpretace (viz KIV/VI), identifikace vhodného pravděpodobnostního modelu (diskrétní/spojité, výběrová rozdělení, případně teoretický model), redukce dimenze (selekce nebo extrakce příznaků)
    2) **zkoumání vztahu více proměnných** – testy závislostí (symetrický vztah proměnných), funkční vztah z hlediska dimenzí (testhy na shody, ANOVA, regresní modely atd.)
    3) **klasifikační metody** – měření vzdáleností pro vícerozměrná data, supervised metody (diskriminační analýza), unsupervised metody (shlukování), hodnocení kvality (train/test split, cross-validation)
  - **aplikace vícerozměrných dat** – modelování náhodných jevů, simulační techniky, zpracování dat, statistická inference

- **vícerozměrná náhodná veličina** – obecný statistický model, stojící za vícerozměrnými daty
  - **definice** – náhodný vektor $\boldsymbol{X}=(X_1,\dots,X_k)$ definovaný na pravděpodobnostním prostoru (Ω, A, P), přičemž $[\boldsymbol{X}\leq\boldsymbol{x}]=\bigcap_{j=1}^{k}[X_j\leq x_j]$
  - **diskrétní náhodný vektor vs. spojitý náhodný vektor**
  - **sdružená distribuční funkce** – reálná funkce $F(\boldsymbol{x})=P(\boldsymbol{X}\leq\boldsymbol{x})$, platí $F(\boldsymbol{x})\in\langle 0,1\rangle$, neklesající, zprava spojitá, $\lim_{x_j\to-\infty} F(\boldsymbol{x})=0$ a $\lim_{x_j\to\infty} F(\boldsymbol{x})=1$
  - **maginální rozdělení** – rozdělení složek náhodného vektoru, tj. jednotlivých atributů,
    - $G(x_1,\dots,x_r)=F(x_1,\dots,x_r,\infty,\dots,\infty)$, diskrétně přes $\sum$, spojitě přes $\int$
    - ze sdružené funkce můžu zjistit marginální, naopak ne
  - **nezávislost náhodných veličin** – sdružená distribuční funkce je rovna součinu marginálních distribučních funkcí
  - **podmíněné rozdělení** – sdružená distribuční funkce děleno marginální podmiňujícího
  - **vektor středních hodnot** – vektor středních hodnot marginálních rozdělení
  - **vektor rozptylu** – vektor rozptylu marginálních rozdělení
  - **kovariance** – stanovena pro všechny páry atributů. statistická míra lineární závislosti
  - **varianční matice** – *var* na diagonále, *cov* všude jinde, symetrická čtvercová matice
  - **kovarianční matice** – pro vektory $\boldsymbol{X}$ a $\boldsymbol{Y}$, není obecně čtvercová, závisí na jejich rozměrech

- **korelace** – různé způsoby definice (Pearson, Spearman, Kendall, ...), $\langle -1, 1 \rangle$
- **korelační matice** – na diagonále 1ky, jinde korelace
- **charakteristiky náhodné veličiny** – střední hodnota, rozptyl, kovariance, korelace
- **charakteristiky podmíněného rozdělení** – střední hodnota, rozptyl, kovariance, korelace
- dále mediány, kvartily atd.

- **přehled vícerozměrných pravděpodobnostních rozdělení**:
  1) **vícerozměrné hypergeometrické rozdělení** – bez náhrady vybíráme vzorky z konečné populace rozdělené do více kategorií
  2) **multinomické rozdělení** – provádíme *n* nezávislých pokusů, každý s *c* možnými kategoriemi výsledku
  3) **dvourozměrné Poissonovo rozdělení** – když modelujeme počty výskytů více různých typů událostí, které mohou být závislé, ale každá má Poissonovu povahu
  4) **dvourozměrné rovnoměrné rozdělení** – stejné pravděpodobnosti => nezajímavé
  5) **dvourozměrná exponenciální rozdělení** – marginální rozdělení mají exponenciální charakter, jsou velmi často používány pro modelování spolehlivosti
     (1) **Gumbelovo** – jednodimenzionální či vícerozměrné rozdělení extrémních hodnot používané pro modelování maxima (nebo minima) náhodných veličin, např. v analýze životnosti nebo extrémních událostí
     (2) **Hougaardovo** – zobecněné rozdělení typu Tweedie, které se používá k modelování nadměrné variability (overdispersion) u Poissonových procesů včetně modelování biologických nebo lékařských dat s variabilní intenzitou
     (3) **Downtonovo** – rozdělení pro dvě Poissonovské veličiny s kladnou závislostí, konstruované pomocí sdílené gamma proměnné, vhodné např. pro modelování závislých počtových dat
     (4) **Arnold-Straussovo** – rozdělení pravděpodobnosti pro modelování závislosti mezi dvěma či více náhodnými veličinami s pevně danou marginální distribucí, často používáno při konstrukci testů hypotéz
     (5) **Freundovo** – dvourozměrné exponenciální rozdělení modelující závislé doby životnosti, kdy selhání jedné komponenty ovlivňuje rozdělení doby selhání druhé
     (6) **Marshall-Olkinovo** – model pro závislé doby přežití, ve kterém mohou komponenty selhat samostatně nebo na základě společné příčiny, běžně používaný v biometrii a spolehlivostní analýze
  6) **vícerozměrné normální rozdělení** – vektor průměrů a kovarianční matice
     – normované normální rozdělení – průměry jsou 0 a kovarianční matice má na diagonále 1ky
     – marginální rozdělení je opět normální rozdělení
     – jinak zas klasicky počítám výběrové charakteristiky
  7) **Wishartovo výběrové rozdělení** – vícerozměrný analog chí-kvadrát rozdělení, používaný k popisu rozdělení kovariančních matic získaných z vícerozměrného normálního rozdělení
  8) **Hotellingovo T^2 býběrové rozdělení** – vícerozměrný analog Studentova t-rozdělení, používaný při testování hypotéz o středních vektorech vícerozměrně normálního rozdělení

- **bodové a intervalové odhady** – funkce výběrových dat, pomocí nichž aproximujeme populační parametr
  - **bodové odhad**y – jedna hodnota, nestranné, konzistentní, efektivní, asymptotická normalizta (CLV); výběrový průměr, výběrový rozptyl, výběrová směrodatná odchylka, SEM atd.
  - **intervalové odhady** – doplněno o nějaké rozmězí, třeba CI u střední hodnoty, chi-kvadrát u rozptylu, pro korelační koeficient pomocí Fisherovy transformace atd.; mohou být dvoustranné/levostranné/pravostranné
  - **(log-)likelihood funkce** – jak dobře určitá hodnota parametru vysvětluje pozorovaná data, skóre kvality pro různé hodnoty parametru (není to pravděpodobnost parametru, ale funkce parametru s pevnými daty), narozdíl od cost function to je pravděpodobnost
  - **skórová funkce** – derivace log-likelihood funkce, Fisherova informační matice (varianční matice skórové funkce)

2

Lineární regrese jedné i více proměnných, odvození cenové/pokutové funkce a techniky její minimalizace, odvození gradientní metody, algoritmus gradientního sestupu, problémy a omezení gradientního sestupu; polynomiální regrese; normální rovnice.

KIV/SU + KMA/MSM

- **lineární regrese jedné proměnné** – z éry před počítači, nejjednodušší supervised regresní algoritmus, moc jednoduchý na dnešní úlohy, nekomplikovaný, step-by-step interpretovatelnost, lze na ni napasovat spoustu úloh a udělat je tak jednodušší k vyřešení
  - **úloha regrese** – aproximace datasetu nějakou funkcí pro inter/extrapolaci dat, obecně nějaká funkce $f(x, \Theta): \mathbb{R}^n \rightarrow \mathbb{R}^k$
  - **hypotéza** – $h_\theta = \theta_0 + \theta_1 x$, což je parametrizovaná lineární funkce (θ_0 je posun na ose y a θ_1 je sklon)
  - **cíl** – najít parametry lineární funkce, které vedou k přímce nejlépe popisující data
  - **state space** (data a přímka) **vs. parametric space** (na každé ose je parametr)
  - **cenová funkce** – ohodnocuje úspěšnost modelu ve vztahu k parametrům, může se maximalizovat nebo minimalizovat, uvádí:

$$J(\theta_0, \theta_1) = arg\,min\left(\frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2\right)$$

  - dělíme na train a test data, abychom zaručili dobrou generalizaci
  - **gradientní sestup** – heuristická optimalizační metoda založená na postupném sestupu ve směru gradientu
    1) **gradient** – vektor parciálních derivací cenové funkce vzhledem k parametrům, typicky musíme použít chain-rule při výpočtu
    2) obecně nemáme zaručeno, že nalezneme globální minimum (můžeme se zaseknout v lokálním)
    3) u lineární regrese jedné proměnné má tvar **paraboloidu** => lokální minimum je i globální minimum
    4) **learning rate** α – zásadní hyperparametr; moc malé (hodně cyklů) vs. moc velké (skáče kolem minima) vs. akorát, zkoumáme pře loss curve
    5) **proces** – inicializace parametrů (náhodně, He, Glorot atd.) => posun parametrů ve směru negativního gradientu o nějaký α násobek => opakovat dokud není splněna nějaká podmínka zastavení (počet iterací, hodnota chyby, změna chyby atd.)

- **lineární regrese více proměnných** – stejný princip, ale pro více více proměnných => maticový zápis:

$$(1) \qquad h_\theta(x_1, \ldots, x_n) = \theta_0 x_0 + \theta_1 x_1 + \ldots + \theta_n x_n \quad | \; x_0 = 1$$

$$(2) \qquad h_\theta(X) = \Theta^T X \quad | \; \Theta = \begin{bmatrix}\theta_0 \\ \theta_1 \\ \vdots \\ \theta_n\end{bmatrix} \in \mathbb{R}^{n+1}, \; X = \begin{bmatrix}x_0 \\ x_1 \\ \vdots \\ x_n\end{bmatrix} \in \mathbb{R}^{n+1}$$

- **komplikuje se parametrický prostor** => komplikuje se chování cenové funkce:
  1) **Rosenbrock funkce** – úžiny kolem kterých se osciluje
  2) **matematické vlastnosti GD** – pomalé kolem optima, derivovatelnost atd.
  3) **feature scaling** – řeší problémy, protože dělá symetričtější funkci kolem minima, často přes normalizaci nebo scaling (nikdy ne na x_0)

- **polynomiální regrese** – komplikovanější funkce, ale princip pořád stejný
  - komplikovanější funkce neznamená lepší model
  - řeším **substitucí** a pak stejný postup jako lineární regrese více proměnných:

$$h_\theta(X) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 \rightarrow h_\theta(X) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3$$

  - tentokrát zcela nutný scaling (mocniny to hodně protáhnou do Rosenbrock)

- **normální rovnice** – analytický přístup řešení lineární regrese více proměnných

$$J(\Theta) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$

$$\Downarrow$$

$$J(\Theta) = \|\Theta X - y\|^2 = (\Theta X - y)^T(\Theta X - y) = \Theta^T X^T X \Theta - 2\Theta^T X^T y + y^T y$$

$$\Downarrow$$

$$\frac{d\,J(\Theta)}{d\,\Theta} = (X^T X)\Theta - X^T y = 0$$

$$(X^T X)\Theta = X^T y \quad \Rightarrow \quad \boldsymbol{\Theta = (X^T X)^{-1} X^T y}$$

  - hledám takové Θ, pro které je gradient roven 0
  - **problémy** – pokud je *n* velké, vektor x_0 musí vždy obsahovat 1ky, X nemusí mít inverzi
  - výhoda, že nemusím řešit parametry

- **robustní alternativy** – když mám outliery, šum atd. (to nahoře v podstatě OLS)
  1) **least trimmed squares (LTS)** – OLD přes několik subsetů dat, vezmeme nejlepší, limitováno kombinatorikou
  2) **least absolute deviations (LDS)** – OLS, akorát místo squares beru absolute values
  3) **M-estimators** – superset metod, obsahuje OLS, MLE, LAD etc., něco minimalizuju
  4) **iteratively re-weighted least squares (IRLS)** – každý estimate je ještě dál vážený
  5) **random sample and consensus (RANSAC)** – vezmu minimum bodů, udělám funkci, vyhodntím, repeat
  6) **Theil-Sen estimator** – křivka jako medián všech párových křivek
  7) **Ridge regression** – když je vysoká kolinearita

- **dodatky z KMA/MSM**:
  - do vzorce regrese se přidává chyba ε, která má typicky normální rozdělení
  - lze dokázat, že pro normální lineární regresi je OLS odhad BLUES
  - **kritéria kvality** – AIC, BIC, R^2, koeficient determinace atd.

3

Logistická regrese, model hypotézy logistické regrese, interpretace výsledků, rozhodovací hranice, klasifikace do více tříd – algoritmus One-vs-All.

KIV/SU

- **binomiální/binární logistická regrese** – predikované hodnoty jsou 0 nebo 1
  - hodnoty mají **Bernoulliho distribuci**
  - **otázka, zda jde skutečně o regresi** – na jednu stranu diskrétní klasifikace, na druhou dochází k fitování sigmoidy podle regresních principů
  - základ pro mnoho jiných klasifikačních ML algoritmů
  - **proč nemůžeme lineární regresi** – hodnoty mimo [0, 1], threshold 0.5 nefunguje (nemusí být dobrý threshold, nemusí dobře separovat, outliery hlavně)
  - **hypotéza** – sigmoida, tj. $h_\theta(X)=\frac{1}{1+e^{-\Theta^T X}}$, speciální case logistické funkce
    1) **interpretace** – pravděpodobnost, že je třída 1, lze vnímat jako podmínenou pravděpodobnost
    2) **rozhodovací hranice** – křivka co odděluje třídy, lineární je že do sigmoidy dám hypotézu lineární regrese, může být ale jakákoliv (třeba radiální)
  - **cenová funkce** – nelze použít tu z lineární regrese, protože vede na non-smooth funkci a tedy GD nefunguje (vede taky na Rosenbrock), proto negative log loss:

$$J(\Theta) = -\frac{1}{m}\left(\sum_{i=1}^{m} y^{(i)}\log\left(h_\Theta(X^{(i)})\right) + (1-y^{(i)})\log\left(1-h_\Theta(X^{(i)})\right)\right)$$

    1) musí se upravit, aby reflektovala možné hodnoty implikované sigmoidou
    2) výsledná forma vychází z maximum likelihood estimation (MLE), konvexní, spojitá, tedy lze GD

- **multinomiální logistická regrese** – predikované hodnoty jsou 0, 1, ..., *K*
  - hodnoty mají binormiální distribuci
  - **one-vs-all** – použiju několik logistických regresí, vždy na jednu třídu vs. zbytek
    1) výslednou třídu pak volím podle toho, který model měl maximální příslušnost

4

Support Vector Machines, cíl optimalizace jako alternativní pohled na logistickou regresi, matematický model SVM, hypotéza s bezpečnostním faktorem, jádra.

KIV/SU

- **SVM** – hledá se optimální nadrovina, která dělá partition dat
  - binární klasifikační metoda, často dobře interpretovatelné a čisté výsledky oproti neuronkám, cheap a na nějaké úlohy zkrátka stačí
- **modifikace logistické regrese** – substituce za $\Theta^T X = z$, pak jde jen o to co je z
  - cenová funkce – stejný princip, zase substituce a jen srovnávání, tedy neg log loss se stane ostrý
  - používá se nějaká konstanta C k regularizaci
  - safety zones – z buď -1 nebo 1, podle toho třída, je to zas ostře
- **cíl optimalizace** – minimalizace kvadratické normy parametrického vektoru θ
- **decision boundary** – optimalizace preferuje nejširší možný margin
  - **kernely** – **lienární** (přímka), **Gaussian** (přes landmarks, k nim se počítá distance, podle toho), **polynomiální** (ne často, horší než kernely), **string** (srovnání stringů), **chi-squared kernel** (na histogramy) atd.
- **parametry SVM** – balancování deviation vs. variance přes C a δ^2 (Gauss)
- **Mercer's theorem** – vlastnost kernelových funkcí, symetrické a pozitivně-definitní
- **multi-class** – zase One vs. All

(hodně matematiky, vše je v přednášce a ve skriptech)

5

Principy testování statistických hypotéz. Testy o shodě středních hodnot, testy o shodě kovariančních struktur.

KMA/MSM

- **testování hypotéz** – proces potvrzení nebo vyvrácení výzkumných otázek statistickým testem
  - není pravda, že nutně prokazuju pravdivost našich tvrzení, je to "jen" statistika
  - typicky vhodné pokud ověřujeme platnost výsledků na úrovni populace, kdy mezi sebou srovnáváme jednotlivé skupiny ze vzorků
  - **druhy statistických hypotéz**:
    1) **parametrická hypotéza** – řeším konkrétní parametr, např. shoda průměrů
    2) **neparametrická hypotéza** – řeším nějakou vlastnosti, třeba typ rozdělení
  - **statistický test** – postup, pomocí ktrého ověřuju mou hypotézu, dle jeho výsledku:
    1) **nulová hypotéza** – většinou nějaká rovnost, nulový efekt, zamítnuta/nezamítnuta
    2) **alternativní hypotéza** – popírá platnost H0, oboustranná vs. jednostranná
  - **hladina významnosti** – typicky se řící Studentovo, Fisherovo nebo chi-square rozdělením, vůči ní řeším výsledek testu
  - rozděleno na kritický obor ve prospěch HA a obor přijetí
  - **druhy chyb**:

		výsledek testu	
		nezamítáme H_0	**zamítáme H_0**
Skutečnost	Platí H_0	Správné rozhodnutí 1 – α (spolehlivost testu)	**Chyba I. druhu** α (hladina významnosti)
	Platí H_A	**Chyba II. druhu** β	Správné rozhodnutí 1 – β (síla testu)

  - **p-value** – pozorovaná hladina významnosti, srovnávám ji s zadnou hladinou; pravděpodobnost chyby I. druhu

  - **testy**:
    1) **shoda střední hodnoty**:
       (1) **Studentův t-test** – jednovýběrový, dvouvýběrový, párový atd.
       (2) **Wilcoxonvů test** – neparam., pro jeden výběr
       (3) **Mann-Whitney U test** – neparam., více výběrů
       (4) **Hotellingův T^2 test** –nD varianta t-testu
       (5) **MANOVA**
    2) **shoda rozptylu**:
       (1) **Fisherův F-test** – param, pro norm
       (2) **Bartlettův test** – heteroskedasticita, normalita
       (3) **Levene/Brown-Forsythe testy** – heteroskedasticita, neparam.

   (4) **Boxův M-test** – srovnání kovar. matic
   (5) **likelihood poměrový test** – zas kovar. matice
3) **normalita** – Shapiro-Wilk, Kolmogorov-Smirnov (s distr. f. norm.), Anderson-Darling, Liliefors atd.
4) **další** – chi-square test dobré shody, ANOVA, MANOVA, ANCOVA, McNemarův test, Fisher's exact test, Dickey-Fuller atd.

(v přednáškách jsou jinak jen definice a rovnice)

6

Snižování dimenze (metoda hlavních komponent - cíle metody, předpoklady metody, odvození a použití metody na příkladech; t-SNE - princip, srovnání s PCA).

KIV/SU + KMA/MSM + KIV/VI

- **redukce dimenze** – cílem je využít korelace/redundance (multikolinearity) mezi daty
  - z pohledu ML **unsupervised learning**
  - cílem je extrakce nejduležitějších dat, snížení dimenze, zjednodušení popisu dat, analýza souvislostí a identifikace outlierů

  - **heuristicko-expertní metoda** – kouknu na atributy a sám určím, co dám pryč
  - **jinak dělení metod**:
    1) **lineární redukce dimenze** – většinou stačí, ale mají limitace
       (1) **PCA založeno na kovarianční matici** – základní a nejčastější
       - **komponenta** – skrytá proměnná, která vysvětluje variabilitu a závislost jednotlivých proměnných, samy o sobě nemají nutně sémanticky smysl
       - **vlastnosti komponent** – $\lambda_1 \geq \lambda_2$ atd., význam složky je podíl λ / suma zbylých λ
       - **vlastnosti báze** – ortogonální, lineárně závislé složky, pokrytí
       - zpětná projekce vede ke ztrátě dat
       - **chceme aby** – komponenty měly klesající význam, nekorelovali, aby nejdůležitější vysvětlila co nejvíc z celkové variability
       - hledá se směr největší variability v datech => soustředíme se na diagonální prvky kovarianční matice
       - **normovat?** – často se dělá, ale pak všude 1ky na diagonále, čímž ztratíme nějaký význam
       - **názvosloví** – komponenty (nově vzniklé atributy), loadingy (koeficienty a_{ij} reprezentující váhové zastoupení původních proměnných), skóre (transformované hodnoty pozorovaných veličin, realizace veličin, souřadnice hodnot v nové bázi)
       - při eigendekompozici hledám vázané extrémy => **Lagrangeova metoda**
       - **vlastní čísla a vlastní vektory** – kovarianční matice je pozitivně semidefinitní a symetrická => existuje jednoznačně eigendekompozice $\Sigma = V \Gamma V^T$, kdy Γ má na diagonále vlastní čísla a V je čtvercová matice vlastních vektorů (ortogonální, tj. $V V^T = I$), použiju SVD
       - **vizualizace PCA** – scree plot, score plot, loading plot, biplot

       (2) **PCA založeno na korelační matici** – v podstatě jako normování předem
       - korelační matice normovaných proměnných je korelační matice nenormovaných proměnných
       - u výběrového souboru je pouze odhad, předpokládáme normalitu

2) **nelineární redukce dimenze** – výrazně náročnější:
   (1) **multidimensional scaling (MDS) –** maximální zachování vzdálenosti mezi datovými body; definuju matici *B* (centered distance matrix) a udělám eigendekompozici => podobně jako PCA, ale ještě s tím dělám dál věci; $O(n^3)$, v praxi rychlé, globální distance
   (2) **t-distributed stochastic neighbor embedding (t-SNE)** – taky hledám body, ale ne přes vzdálenost klasickou, ale stochastickou přes Kulback-Leiblerovu divergenci (KLD); podobnost definována jako podmíněná pravděpodobnost, že si point *i* bezme *j* jako souseda; hyperparametr perplexity (neighbor size); iterativní; $O(n^2)$ v praxi pomalé, stochastické, dobré na local structure
   (3) **uniform manifold approximation and projection (UMAP)** – rychlejší, buduje se grafová struktura sousedství a snahou je ji nejvíce zachovat
   (4) **encoder-decoder neuronky** – smrsknu data a roztáhnu

7

Shluková analýza, cíle metody, předpoklady metody, odvození a použití metody na příkladech. Optimalizační kritérium K-means, výběr centroidů, volba počtu shluků.

KIV/SU + KMA/MSM

- **shluková analýza** – cílem je vytvořit shluky tak, aby si prvky uvnitř shluků byly co nejvíce podobné a skupiny mezi sebou co nejvíce odlišné
  - **unsupervised learning**, typicky použito u **data miningu**
  - **kroky** – volba metriky a algoritmu, realizace algoritmu, zhodnocení kvality
  - **typy shlukové analýzy** – podle několika kritérií:
    1) **zda je nebo není hierarchické**:
       1) **hierarchické** – lze vizualizovat jako dendrogramy
          (1) **aglomerativní** – shluky vznikají spojováním (bottom-up)
          (2) **divizní** – dělím shluky na podshluky (top-down)
       2) **nehiearchické** – určíme počet shluků a pak se to spočte:
          (1) **k-means** – viz níže
          (2) **MacQueenova metoda** – iterativní verze k-means, která aktualizuje centroidy po každém přiřazení bodu do shluku
          (3) **Wishartova metoda RELOC** – shlukovací algoritmus založený na hustotě, který umožňuje relokaci bodů mezi shluky podle lokální hustoty a pravděpodobnostních kritérií
          (4) **metoda ISODATA** – vylepšený k-means algoritmus, který dynamicky mění počet shluků jejich dělením a slučováním podle statistických kritérií
    b) **podle počtu shluků**:
       1) **striktní shlukování** – každý objekt jednomu shluku
       2) **striktní shlukování s outliery** – každý objekt jednomu nebo žádnému
       3) **překrývající se shluky** – objekt více shlukům zároveň
       4) **hierarchické shluky** – shluky mají charakter rodič-potomek
       5) **subspace shluky** – hledání clusterů v různých subspaces datasetu
    c) **podle reprezentace**:
       1) **hard clustering** – každý sample jen jednomu, reprezentováno přes nějakou diskrétní příslušnost 1/0
       2) **soft clustering** – každý sample více, reprezentováno přes nějakou míru příslušnosti nebo pravděpodobnost příslušnosti
    d) **podle typu modelu**:
       1) **connectivity models** – např. hierarchické shlukování, protože tvoří shluky na základě distance connectivity
       2) **centroid models** – např. k-means, protože reprezentuje každý cluster centroidem
       3) **distribution models** – např. GMM, používají se nějaké multivariate statistické distribuce

4) **density models** – např. DBSCAN a OPTICS, clustery jsou dense regiony v data space
5) **subspace models** – např. Biclustering, clustery jsou modelovány oběma clustery a relevantními atributy
6) **group models** – neposkytují refined model for their results a jen poskytnou groupin informaci
7) **graph-based models** – např. HCS, ideální prototyp je hledání klik, ale můžu i kvazi-kliky

- **techniky shlukování obecně**:
  1) **Q-techniky** – hledám vzdálenost mezi objekty (řádky), metriky např. Eukleidovská, Manhattan, Minkowski, Lagrange, Canberry, Mahalanobis
  2) **R-techniky** – hledám vzdálenost mezi atributy (sloupci), metrika často nějak založená na korelaci
- **využití** – marketing, sales, plánování výroby, cybersecurity, research atd.

- **k-means** – *n* pozorování do *k* clusterů, každá tomu clusteru, k jehož centroidu má nejblíže
  - vždy stabilní, nikdy nediverguje (teoreticky, prakticky může dojít k oscilaci kvůli přesnosti desetinné čárky
  - **centroid** – reprezentuje celý shluk
  - **2 kroky, které jsou iterovány**:
    1) **tvorba clusterů** – každý sample je přiřazen podle blízkosti centroidů (ty na začátku nějak inicializovány, random nebo přímo z dat)
    2) **přesun clusterů** – centroid se přesune do průměrné lokace přiřazených bodů
  - algoritmus se docela často zasekne v lokálním minimu => **dělám několik inicializací a beru tu nejlepší**
  - **cenová funkce** – pro algoritmus není potřeba, ale můžůeme podle ní hodnotit kvalitu a pak dál volit z různých variant, např. MSE/MAE atd. mezi body a nejbližším centroidem
  - **volba počtu shluků**:
    1) **manuálně** – můžu otestovat několik variant a vybrat tu nejlepší třeba podle nějakého indexu kvality (např. Silhouette Score, Davies-Bouldin index, Calinski–Harabasz index atd.)
    2) **elbow metoda** – udělám víc variant a hledám, kde dojde k největšímu zlomu v cenové funkci (ta vždy klesá)
    3) **mám dáno** – nejlepší varianta, rovnou vím co hledám

- **Gaussian mixture models (GMM)** – směs gaussovských rozdělení, kterých je *k*
  - soft, každé GMM definováno (π, μ, Σ), suma vah musí být 1
  - nelze použít MLE => používáme **EM algoristmus**:
    1) **E-step** – pro současný model estimate spočti membership probs.
    2) **M-step** – přiřadit podle max probs. => přesunu parametry podle dat
  - **membership probability**:

$$p(z_k = 1|\boldsymbol{x}^{(n)}) = \frac{\pi_k \mathcal{N}\left(\boldsymbol{x}^{(n)}|\mu_k, \boldsymbol{\Sigma}_k\right)}{\sum_{j=1}^{K} \pi_j \mathcal{N}\left(\boldsymbol{x}^{(n)}|\mu_j, \boldsymbol{\Sigma}_j\right)}$$

8

Diskriminační analýza, cíle metody, odvození a použití metody na příkladech.

KMA/MSM + KIV/SU

- **diskriminační analýza** – hledáme rozhodovací pravidlo, které na základě hodnot náhodného vektoru rozdělí do několika tříd => supervised klasifikace, pokud znám rozdělení z apriorní informace odjinud
- **složky diskriminační analýzy**:
  1) na train množině sestavím rozhodovací pravidlo (tzv. diskriminační část)
  2) zhodnotím kvalitu rozhodovacího pravidla z pohledu trénovací množiny
  3) zhodnotím kvalitu rozhodovacího pravidla z pohledu validační množiny
- mám náhodný *k*-rozměrný vektor *X* s funkcí hustoty $f_k(x)$, A_j je náhodný diskrétní jev vyjadřující příslušnost *X* k *j*-té třídě, $P(A_j)=\pi_j$ je aprorní pravděpodobnost, $S\in\mathbb{R}^k$ je nějaký prostor
  - cílem je *S* rozdělit na S_j **disjunktních oblastí**, tj. $S=\bigcup_{i=1}^{p} S_j$ a $S_j\cap S_i=0$ pro $i\neq j$
  - rozhodovací pravidlo založené na **maximum likelihood** => předpokládáme znalost rozdělení dílčích A_j: $L_j(x)=g_j(x)=max\, g_i(x)$ pro $i=1,2,\dots,p$
  - **matematicky je rozhodnutí**: $S_j=\{x: L_j(x)>L_i(x) \text{ for } i=1,2,\dots,p; i\neq j\}$

- **obecná lineární diskriminační funkce (LDF)** – pokud $A_j\sim N_k(\mu_j,\Sigma)$, pak je maximum likelihood ekvivalentní minimalizaci kvadrátu Mahalanobisovy vzdálenosti mezi *x* a μ_j: $\theta^2(\boldsymbol{x},\mu_j)=(\boldsymbol{x}-\mu_j)^T\Sigma^{-1}(\boldsymbol{x}-\mu_j)$, pokud *p* = 2, pak: $S_1=\{x;\alpha^T(\boldsymbol{x}-\mu)\geq 0\}$
  - **Mahalanobisova vzdálenost** – míra vzdálenosti mezi bodem a rozdělením nebo mezi dvěma body, která bere v úvahu korelace mezi proměnnými a měřítka jednotlivých proměnných

- **kvadratická disktiminační analýza (QDA)** – neuvažuje shodné kovarianční matice, v diskriminační funkci je kvadratický člen

- **Bayesovo rozhodovací pravidlo** – navíc používám zmíněný prior:

$$\mathcal{P}(A_j|x)=\frac{\mathcal{P}(A_j)\cdot f_j(x)}{\sum_{j=1}^{p}\mathcal{P}(A_j)\cdot f_j(x)}=\frac{\pi_j\cdot f_j(x)}{\sum_{j=1}^{p}\pi_j\cdot f_j(x)}$$

  - množina S_j dána tvarem: $S_j=\{\boldsymbol{x},\pi_j f_j(x)\geq\pi_i f_i(x);\ \ i=1,2,\dots,p;\ \ i\neq j\}$
  - **LDF pro Bayesovo pravidlo** – pokud předpokládáme shodnou variační strukturu, je nerovnost jednodušší: $L(x)=\beta^T\boldsymbol{x}+\gamma>0$

- **ekonomické ocenění rozhodovacího pravidla** – pracuji s pravděpodobností chybného zařazení, např. p_{12} je chybné zařazení *i* do *j*: $p_{ij}=P(x\in S_i|A_j)=\int_{S_i} f_j(x)\,dx$

- odpovídající ocenění c_1 a c_2 použijeme ro **expected cost of misclassification (ECM):** $ECM = c_{21} p_{21} \pi_1 + c_{12} p_{12} \pi_2$
- minimalizujeme a získámé množinu $S_1 = \left\{ x ; \frac{f_1(x)}{f_2(x)} \geq \frac{c_{12} \pi_2}{c_{21} \pi_1} \right\}$
- c_{ij} je celková ztráta, střední hodnota ztráty prvků A_j je $c_j = \sum_{i=1}^{p} c_{ij} p_{ij}$, střední hodnota celkové ztráty je $c = \sum_{j=1}^{p} c_j \pi_j$, hledáme disktiminační skóre
- minimalizace ztráty je ekvivalentní maximalizací diskriminačního skóre

- **Fisherova diskriminační analýza** – hledám lineární transformaci pro X, při které minimalizuju OLS odchylek mezi třídami ku OLS uvnitř tříd
  - hledáme $a^* \in \mathbb{R}^k$, aby: $a^* = argmax \frac{Q_B}{Q_W} = argmax_{a \in \mathbb{R}^k} \frac{a^T Ba}{a^T Wa}$, což je vlastní vektor příslušící největšímu vlastnímu číslu matice $W^{-1} B$

- **hodnocení kvality**:
  1) **confusion matrix** – ukazuje TP, TN atd.
  2) **train/test split** – standard v ML
  3) **cross-validation** – postupně vynechávám různé části a testuju
  4) **apparent error rate (APER):** $\frac{\sum_{i \neq j} n_{ij}}{n}$
  5) **actual error rate (AER):** $\frac{n_{12} + n_{21}}{n_1 + n_2}$
  6) **reciever operating characteristics (ROC)** – hodnocení senzitivity a specificity
  7) **area under curve (AUC)** – plocha pod ROC křivkou
  8) **Wilcoxonova lambda** – pro hodnocení Fisherovy diskriminační analýzy: $\Lambda = \frac{det(W)}{det(T)} = \frac{det(W)}{det(W+B)}$

9

Různé možnosti vizualizace kvantitativních dat: scatter ploty, histogramy, boxploty, violin ploty, paralelní souřadnice, ….

KIV/VI

- **scatter plot** – body v prostoru, můžu kódovat barvy, velikost atd.
  - **scatter plot matrix (SPLOM)** – u nD dat, kdy udělám matici scatterplotů
- **heatmapa** – buňce je přiřazena barva, např. korelační matice, matice sousednosti
  - lze řešit přeorganizování pořadí, např. přes cosinovou vzdálenost, UPGMA atd.
- **radar chart** – linie do kruhu, v polárních souřadnicích atributy
- **liniový graf** – klasický liniový graf
- **sloupcový graf** – co sloupec to nějaká diskrétní jednotka (typicky kategorie), výška sloupce odpovídá hodnotě
- **histogram** – viz otázka 11
- **pie chart** – podíl výseče odpovídá celkovému podílu, musí dát dohromady 100 %
- **boxplot** – viz otázka 11
- **violin plot** – viz otázka 11
- **paralelní souřadnice** – osa *x* nemá význam, několik os *y* a instance je nějaká lomená čára mezi nimi
  - **slope graf** – speciální případ když mám 2 osy
  - lze řešit přeuspořádání tak, aby bylo co nejvíc vidět
  - **semitransparentnost** – pokud se překrývá hodně instancí
  - **polylinie** – na každé axis KDE, pak bundles, pak spojit přes Beziérovi křivky
  - **3D** – několik kolem jedné, místo os roviny
- **RadViz** – jakoby napnuté body na pružinách, lze řešit clustering
  - **RadViz Deluxe** – automaticky nastaví atributy na kružnici
- **infografika** – jednoduchý a efektivní nástroj, hodně faktů na malé ploše, má zaujmout čtenáře, často hodně grafické, vhodné pro explanatory vizualizaci komplexních témat veřejnosti
- **mapa** – přidává geografickou prostorovou informaci, např. kartogramy, kartodiagramy, dasymetrická metoda, anamorfovaná mapa atd.
- **nejistota** – viz otázka 11
- **barvy** – viz otázka 11

- **ekonomická data** – specifické metody:
  1) **moving average + Bollinger bands** – trend a okolo nějaká deviace
  2) **candlestick chart** – open/close/low/mid, bullish (zelené, open dole) vs. bearish (červené, open nahoře)
  3) **OHLC chart** – alternativa k candlestick
  4) **Heiken-Ashi chart** – open/close mají jiný význam (průměry za minulé a současné období), low/high je pak min/max O,C,L/O,C,H
  5) **Kagi chart** – časová osa není lineární, řeší jen změny

6) **point and figure chart** – supply/demand vztah, řeší počet buyers a sellers křížkama a kolečkama na čtvercové mřížce
7) **Sankey diagram** – flow, třeba peněz mezi institucemi
8) **flowmap** – v podstatě Sankey, ale na mapě v prostoru

- **časové řady** – specifické metody:
  1) **Ganttův diagram** – management projektu, paralelní procesy
     - **Planninglines** – v podstatě fuzzy Gantt
  2) **spiral graph** – data do spirály
  3) **cycle plot** – složky sezónnosti i trendu
  4) **lag plot** – scatterplot hodnoty a hodnoty o něco v minulosti
  5) **Pertův diagram** – v podstatě kritická cesta
  6) **time wheel** – čas uprostřed, atributy okolo
  7) **streamgraph** – baseline se hýbe, lze číst jen poměry, ale ne absolutní hodnoty

10

Vizualizaci hierarchií a relací: node-link diagrams, techniky rozložení uzlů (layout), containment diagrams (např. Treemap), layering (např. icicle plots); maticová vizualizace rozsáhlých grafů.

KIV/VI

- **graf** – uzly a hrany, oboje má své vlastnosti, sledujeme vztahy
- **hierarchie** – acyklický graf sestavený podle důležitosti uzlů
  - **strom** – hierarchie co má jeden root, každý strom je hierarchie, ale ne každá hierarchie je strom

- **metody vizualizace hierarchií**:
  1) **indentation** – O(n), snadné, styl vizualizace jako directory tree
  2) **node-link diagrams** – uzly spojeny hranami
     - **varianty** – horizontální, vertikální, radiální
     - **Reingold-Tilford algoritmus (1981)** – layout (pouze) binárních stromů, O(n)
     - **Walker (1990)** – obecné stromy, $O(n^2)$
     - **van der Ploeg (2013)** – i pro non-layered nodes
  3) **containment diagrams** – uzly jsou tvary, které obsahují subtrees, velikost určena velikostí subtree a hodnotou uzlu
     (1) **circle-packing layout** – kruhy, spousta prázdného místa, heuristiky, ale vidím docela jasně
     (2) **treemap** – nějaký prostor rekurzivně dělím podle poměru hodnot uzlu, často do obdélníku nebo čtverce
     (3) **squarified treemap** – řeší stav, kdy mám hodně proužků, zachová nějaký daný poměr, ztratím ale víc info o hierarchii
     (4) **Voronoi treemaps** – stejný princip ale Voronoi polygony, O(n log n)
  4) **layering** – dělím postupně na subspaces, obsahuje prázdná místa
     (1) **icicle plot** – horizontální nebo vertikální
     (2) **sunburst chart** – radiální
  5) **další** – RIT, VineMap, Hi-Tree, BaobabView atd.

- **metody vizualizace grafů**:
  1) **node-link diagramy** – zobecnění toho samého co u hierarchií
     (1) **arc diagram** – 1D, spojeno oblouky, nutné řešit pořadí uzlů
     (2) **chord diagram** – radiální, ideálně zas nějak ordered
     (3) **grid layout** – blbé, bokud použiju pravoúhlé hrany, jinak asi OK
     (4) **force-directed** – analogie k fyzikálnímmu pnutí podle hodnot hran a uzlů, snaha najít nějaké equilibrium
  2) **matice sousednosti** – umožňuje pozorovat vzory, identifikovat kliky, clustery atd., může být až moc velká
- **bundling** – když je hodně hran, můžu sloučit (můžu mergovat i uzly)

- pokud je zajištěna **interaktivita**, může být docela dobré ve 3D
  - často nějaká navigace, select a drag
  - **lenses** – jakoby lokální zoom rybím okem na nějaký uzel
  - často details on demand
- vše lze všelijak barvit, kódovat věci jako velikost, sílu vztahu apod.
- **blbé stavy vizualizací**:
  1) **hairball** – celkově chaos
  2) **snowstorm** – spousta malých grafů
  3) **starburst** – jeden centrální a spousta okolních uzlů

- **security** – grafy jsou záadní, často velmi velké
  - **predicate node** – uzel co má sémantickou vlastnost, např. všechny IP adresy z webu
  - ze vzorů chord diagramu a Sankey diagramu lze vyhodnotit útoky (DDoS, ping atd.)

- **text** – taky některé metody:
  1) **TextArc** – konkordance slov uspořádaná do spirály
  2) **WordTree** – v podstatě hierarchie podle váhy a četností
  3) **Phrase Net** – v podstatě wordcloud pospojovaný podle konkordance
  4) **EmojiText (2021)** – sentiment analysis

11

Vizuální manipulace - konkrétní příklady a jejich řešení; volba barevné škály, vizualizace nejistoty.

KIV/VI

- **vizuální manipulace** – úmyslné (nebo neúmyslné) zkreslení informace pomocí grafických prostředků
  - u vizualizací nepracujeme jen s informací jako takovou, ale i její percepcí => zcela korektní data vizualizovaná na správném grafu, ale třeba jen špatnými barvami, lze vnímat jako formu vizuální manipulace
  - **typické metody manipulace**:
    1) **mlžení v titulu** – pojmenování grafu zavádějícím způsobem, který vede k misinterpretaci (např. Bush vs. Obama)
    2) **zkrácená osa *y*** – typicky u sloupcových grafů, kde vede k zvýšenému kontrastu mezi sloupci, případně časté při "dokazování bankrotu", kdy se ukáže nějaký lokální výkyv přes celý graf
    3) **zkrácená osa *x*** – typicky u časových řad, kdy lze zvolit pouze určité časové okno, které má za cíl něco prosadit (např. že firma klesá, přestože v globálu stoupá)
    4) **nepoměrné rozměry** – typicky když počet kóduju do plochy, např. místo sloupcového grafu udělám množstsví velikostí čtverce
    5) **nevhodný rozměr grafu** – grafy vypadají percepčně jinak placaté nebo naopak vysoké, ideál něco mezi nebo přímo 1:1 když lze
    6) **několik os** – lze vytvořit představu, že spolu nějaké jevy souvisí (např. že se kříží)
    7) **3D grafy** – zkreslená interpretace, typicky pie chart, kdy je zájmová skupina v popředí (výseče dál se zdají menší)
    8) **nevhodné pořadí** – typicky u sloupcového grafu, kdy z něj nelze jednoznačně usoudit, kdo má větší nebo menší hodnotu
    9) **zavádějící volba barev** – zneužití zaběhlých norem, např. zobrazit zeleně negativní jev a červeně pozitivní
    10) **specifické u map** – volba kartografické projekce (každá něco nějak zkresluje, takže můžu např. udělat nějaké území mnohem větší, než ve skutečnosti je, což se hodí např. u politické propagandy) a přehnaná generalizace (můžou zmizet důležité objekty)

- **barevná škála** – zásadní při vizualizaci téměř všech typů dat
  - **kvalitativní data** – barevná škála užívá odstín pro odlišení
  - **kvantitativní data** – barevná škála užívá jas a saturaci pro odlišení
    1) **diskrétní (tj. ordinální)** – gradient s diskrétním odstupňováním
    2) **spojitá** – gradient s diskrétním nebo spojitým odstupňováním
       (1) přestože je spojitá stupnice fakticky správnější, je z hlediska percepce stejně lepší použít diskretizovanou do max. **5–7 barev** (lze pak skutečně určit hodnotu z grafu, u spojitých barev to může být velmi těžké)

    (2) **rozdělení přes** – **rovnoměrné intervaly** (vše stejně), **kvantily** (podle počtu), **natural breaks** (trošku něco jako segmentace), **statistické charakteristiky** (třeba odchylky apod.), **manuální** (většinou nutné jako finální krok)
  - **unipolární** (sekvenční, tj. jen do jedné barvy) **vs. bipolární** (divergentní, tj. do dvou barev, když mám 0 nebo třeba průměr)
  - **bivariantní stupnice** – snaha spojit dva atributy do jedné stupnice, riskantní a může se hodně blbě interpretovat
  - **duhová stupnice** – příšerná věc, prakticky na nic se nehodí, ale přesto se široce používá
  - **percepce barev se mění** – podle kultury, okolních barev, velikosti obarvené plochy, barvosleposti, vzdělání, věku, kontextu vizualizace atd.

- **vizualizace nejistoty** – drtivá většina datasetů nemá data za populaci, ale nějaký sample => děláme nějaké odhady, takže je nutně i nějaká nejistota
  - **precision** (závislá na rozptylu) **vs. accuracy** (závislá na bias)
  - **zdroje náhodných chyb** – chyba v měření, přirozená variabilita, špatně navržený experiment atd. (chyba klesá s počtem vzorků)
  - **zdroje systematických chyb** – **sampling bias** (dotazník vyplní určití lidé), **nonresponse bias** (ti co se neúčastní nejsou zohledněni), **social desirability bias** (odpovídají to co chceme slyšet, ne co si myslí) atd.
  - **obecné přístupy k vizualizaci nejistoty**:
    1) **kompletní vizualizace** – nic se nezatajuje, ale rychle je graficky nečitelné (pomůže jitter nebo semitransparentní barvy)
    2) **agregace** – typicky průměr, mediín, min/max atd., problém s outliery, může skrýt některé důležité vzory
  - průměr jen u dat s normálním rozdělením
  - **konfidenční interval (CI)** – jsme si na X % jistí, že populační parametr je v daném rozsahu (tj. pokud udělám několik samples z populace, tak X % z nich bude mít sample mean v tomto intervalu)
  - **metody vizualizace nejistoty**:
    1) **error bars** – dávají se k sloupcovým grafům, typicky SD, SE nebo 95% CI
       (1) docela překvapivě s nimi má spousta lidí problém a neví jak je interpretovat
    2) **"point with error bars"** – dělá menší problémy než sloupce
    3) **box plot (také Tukey box nebo Box and whiskers plot)** – velmi častá vizualizace, která může nést mnoho významů
       (1) 95% CI a 50% CI
       (2) min/max a kvartily (+ outliery, které jsou 3/2 násobek hodního/dolního kvartilu)
       (3) lze odstrani komplet box, což je trochu vizuálně čistší
       (4) boxplot + jednotlivé body (protože může pořád skrýt strukturu)
    4) **vase plot** – box je nahrazen symetrickým odhadem hustoty
    5) **gradient plot** – rozmažu okraje (dost dobře interpretováno)
    6) **violin plot** – vykreslení distribučních funkcí (+ často ještě uvnitř boxplot), lze dobře dělit na dvě kategorie
    7) **quantile regression** – isokřivka procházející daným kvantilem, často také s CIs
    8) **histogram** – počet instancí ve vymezených binech
       (1) kumulativní vs. nekumulativní

(2) 1D vs 2D (vs. 3D?

9) **bivariate škály** – s klesající jistotou klesá intenzita barvy, těžko se interpretuje, když už tak ubírat s nejistotou barvy (pak stupnice kruhová výseč)
10) **v mapách** – složité, buď přes glyfy nebo šrafování

- u vizualizací nepracujeme jen s informací jako takovou, ale i její percepcí => zcela korektní data vizualizovaná na správném grafu, ale třeba jen špatnými barvami, lze vnímat jako formu vizuální manipulace
- liniové grafy lze vizualizovat obdobně